



## Investigations of oligonucleotide usage variance within and between prokaryotes

Bohlin, J.; Skjerve, E.; Ussery, David

*Published in:*  
P L o S Computational Biology (Online)

*Link to article, DOI:*  
[10.1371/journal.pcbi.1000057](https://doi.org/10.1371/journal.pcbi.1000057)

*Publication date:*  
2008

*Document Version*  
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

*Citation (APA):*  
Bohlin, J., Skjerve, E., & Ussery, D. (2008). Investigations of oligonucleotide usage variance within and between prokaryotes. *P L o S Computational Biology (Online)*, 4(4), e1000057.  
<https://doi.org/10.1371/journal.pcbi.1000057>

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Investigations of Oligonucleotide Usage Variance Within and Between Prokaryotes

Jon Bohlin<sup>1</sup>, Eystein Skjerve<sup>1</sup>, David W. Ussery<sup>2\*</sup>

<sup>1</sup> Norwegian School of Veterinary Science, Oslo, Norway, <sup>2</sup> Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark

## Abstract

Oligonucleotide usage in archaeal and bacterial genomes can be linked to a number of properties, including codon usage (trinucleotides), DNA base-stacking energy (dinucleotides), and DNA structural conformation (di- to tetranucleotides). We wanted to assess the statistical information potential of different DNA 'word-sizes' and explore how oligonucleotide frequencies differ in coding and non-coding regions. In addition, we used oligonucleotide frequencies to investigate DNA composition and how DNA sequence patterns change within and between prokaryotic organisms. Among the results found was that prokaryotic chromosomes can be described by hexanucleotide frequencies, suggesting that prokaryotic DNA is predominantly short range correlated, i.e., information in prokaryotic genomes is encoded in short oligonucleotides. Oligonucleotide usage varied more within AT-rich and host-associated genomes than in GC-rich and free-living genomes, and this variation was mainly located in non-coding regions. Bias (selectional pressure) in tetranucleotide usage correlated with GC content, and coding regions were more biased than non-coding regions. Non-coding regions were also found to be approximately 5.5% more AT-rich than coding regions, on average, in the 402 chromosomes examined. Pronounced DNA compositional differences were found both within and between AT-rich and GC-rich genomes. GC-rich genomes were more similar and biased in terms of tetranucleotide usage in non-coding regions than AT-rich genomes. The differences found between AT-rich and GC-rich genomes may possibly be attributed to lifestyle, since tetranucleotide usage within host-associated bacteria was, on average, more dissimilar and less biased than free-living archaea and bacteria.

**Citation:** Bohlin J, Skjerve E, Ussery DW (2008) Investigations of Oligonucleotide Usage Variance Within and Between Prokaryotes. *PLoS Comput Biol* 4(4): e1000057. doi:10.1371/journal.pcbi.1000057

**Editor:** Gary Stormo, Washington University, United States of America

**Received:** October 12, 2007; **Accepted:** March 12, 2008; **Published:** April 18, 2008

**Copyright:** © 2008 Bohlin et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** Part of this work was funded by the Danish Center for Scientific Computing, DU is funded by grants from the Danish Research Council, and JB is funded by the Norwegian School of Veterinary Science.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: dave@cbs.dtu.dk

## Introduction

Prokaryotic DNA can be considered as a long chain of overlapping oligonucleotides, and frequencies of differently sized oligonucleotides can reveal different properties and patterns of bacterial and archaeal genomes. On average, roughly 86% of prokaryotic DNA codes for proteins [1] and thus a considerable amount of information is held in trinucleotide (codon) frequencies. Additional information however, can be found by studying other oligonucleotide sizes. Dinucleotide distributions, or nearest neighbor frequencies, are used to calculate base-stacking energies [2], while DNA structural properties can be calculated using di- to tetranucleotide frequencies [2–5]. In addition, the structures of A, B and Z type DNA helices are largely determined by 11-, 10- and 12-mers, respectively [2,6]. Another advantage of considering genomic DNA as a set of fixed-sized oligonucleotide frequencies is that bias and pattern preference, i.e. the randomness inherent or lack thereof, in the complete DNA sequence can be detected [3]. Alternatively, DNA patterns can be investigated by examining occurrences of individual nucleotides [7,8].

Oligonucleotide frequencies are very much influenced by codon distributions which, in turn, are correlated with GC-content [9]. Since GC content is correlated with the environment of prokaryotes [9,10], so are oligonucleotide distributions [11]. The distributions of oligonucleotide frequencies can reveal other

properties as well. GC skews, i.e. increased cytosine compared with guanine content on the leading strand, can be used to determine DNA replication start and stop positions in bacteria [12]. However, many archaeal and slow replicating bacterial genomes do not have pronounced GC skews (or AT skews) on leading and lagging strands, but replication start and stop positions can be detected with increased precision by examining oligonucleotide frequency skews with progressively larger oligonucleotide sizes [13].

In addition to the properties described above, transcription and regulation sites are also coded by certain oligonucleotide patterns [14]. Such oligonucleotides are therefore thought to be severely under- or overrepresented compared with what is expected from mean genomic GC content or compared to oligonucleotide frequencies found in other closely related species [14].

The examples above illustrate some of the properties that can be extracted from DNA sequences by examining oligonucleotide usage variance. This motivated us to explore how oligonucleotide distributions change within and between prokaryotes in coding and non-coding regions, how biased oligonucleotide frequencies are, and whether any particular trends could be detected. In order to do this a series of statistical tests were performed on all sequenced bacterial and archaeal chromosomes (up to September 2006). We found that tetranucleotide frequencies carried considerable genomic information potential, and were therefore used in

## Author Summary

There are potentially many factors responsible for how archaeal and bacterial genomes are composed. Recent advances in DNA sequencing have made it possible to use computational and statistical methods to examine the interplay between evolution and genomic composition. We wished to see whether particular properties could be extracted that would provide clues on how prokaryotic DNA is composed. For instance, we wondered whether or not protein coding regions carried a greater information potential than non-coding regions, if there is a link between genome size and GC content, whether GC content is different in coding and non-coding regions, and possible associations between DNA composition and environment. Our results indicated that genomic nucleotide frequencies are a determinant of many DNA compositional properties, but also that other influences are at work. For instance, bacteria are known to frequently exchange DNA with the environment and other organisms. Acquired DNA can therefore have different compositional properties than host DNA, and since pathogenicity and antibiotic resistance in bacteria is often associated with foreign DNA, advancing the knowledge of DNA composition is of great importance.

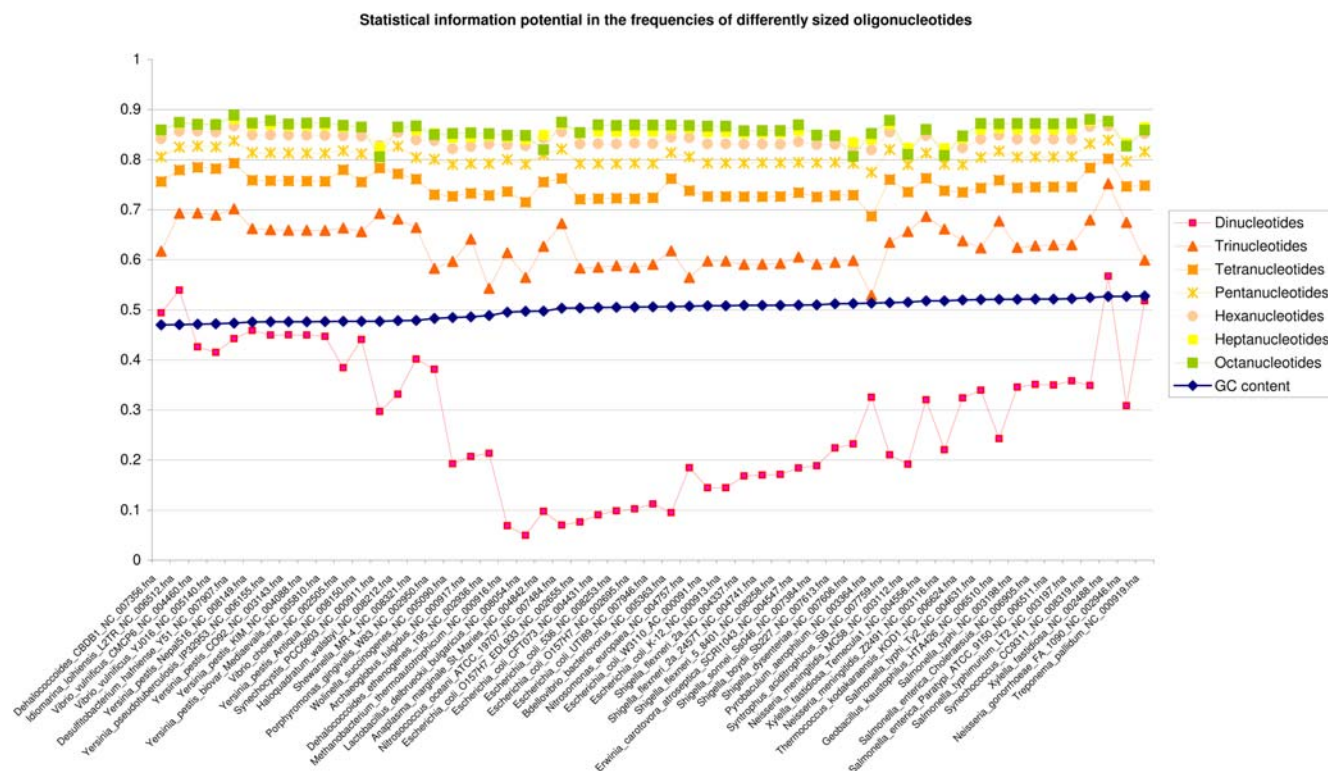
all statistical tests based on oligonucleotide usage. The statistical tests included oligonucleotide usage deviation from mean (OUD, a measure of oligonucleotide frequency variations in genomes), and oligonucleotide usage variance from expected (OUV, a measure of

randomness or bias [3]). These tests were used to examine how tetranucleotide frequencies fluctuated within chromosomes (OUD), and how tetranucleotide distributions differed from random tetranucleotide frequencies calculated by mean genomic nucleotide frequencies (OUV). OUV measures GC content both globally and locally in chromosomes with a 40 kbp non-overlapping sliding window and was used to calculate expected tetranucleotide frequencies as well as GC content in coding and non-coding sequences for each chromosome tested. These tests were performed for all sequenced 402 prokaryotic chromosomes at the time, and their corresponding open reading frames. The first test, however, was concerned with statistical information potential in different oligonucleotide sizes using a different approach than [3].

## Results/Discussion

### Information Potential in Oligonucleotides

We measured the statistical information carried by the differently sized oligonucleotides from di- to octanucleotides in prokaryotes with GC contents between 47% and 53%. From Figure 1, it can be observed that the largest increase in information was obtained by going from nucleotide frequency approximation of dinucleotides to trinucleotide usage approximations based on dinucleotide frequencies and GC content (details can be found in Materials and Methods). A more careful investigation of Figure 1 revealed that progressively less information was gained from usage approximations of tetranucleotides up to heptanucleotides, and practically no additional information appeared to be present in



**Figure 1. Statistical information potential in differently sized oligonucleotides.** Cumulative information potential is measured in di- to octanucleotide frequencies in prokaryotic genomes with GC content between 47% and 53%. These genomes were selected because of the increased sensitivity of the Pearson correlation measure for chromosomes with similar AT/GC content. The archaeal and bacterial chromosomes are represented along the horizontal axis, sorted by increasing GC content from left to right, with corresponding correlation scores between observed  $n$ -mer words and approximated  $n$ -mer words on the vertical axis. The  $n$ -mer words were approximated by observed  $(n-1)$ -mer words and genomic nucleotide frequencies. High correlation scores indicate increased similarity between observed and approximated oligonucleotide usage.  
doi:10.1371/journal.pcbi.1000057.g001

approximated octanucleotide frequencies. Thus, oligonucleotide sizes larger than hexanucleotides possess little additional information potential, if any, in prokaryotic DNA.

### GC Content in Coding and Non-Coding Regions

GC content was measured in coding and non-coding regions (see Materials and Methods for more detail) and it was found that non-coding regions were, on average, roughly ~5.5% more AT rich (~5.3% and ~5.5%, for AT and GC rich chromosomes respectively) than coding regions, with the assumption that 14%, on average, of each chromosome was non-coding DNA [1].

Why non-coding regions are ~5.5% more AT rich than coding regions may be related to the DNA curvature in promoter regions, and possibly termination sites [15], and to lower stacking energies found in AT rich DNA patterns compared with GC rich [2]. Increased AT content in non-coding regions suggest less energy is required to split the double helix for transcription [2].

### GC Content and Genome Size

Although the link between GC content and genome size has been debated [16,17], we obtained significant ( $P < 0.001$ ) correlation with  $\rho = 0.47$  (Spearman's rho) between chromosome size and GC content (see Figure 2). The following regression equation was fitted:

$$Y_{size} = \exp(0.15 + 1.92X_{GC})$$

with the assumption of linear variance.  $Y_{size}$  gives the size of the chromosomes (response) in mbp and  $X_{GC}$  is global GC content (predictor,  $P < 0.001$ ).

### Bias in Tetranucleotide Usage

We measured how tetranucleotide usage varied in genomes compared with expected tetranucleotide usage. This expected tetranucleotide usage was calculated from mean genomic GC content, and implicitly assumes that each nucleotide in every tetranucleotide, and therefore also the whole chromosome, is independent of its neighbors. In other words, the more similar observed and expected tetranucleotide frequencies are, the more random (i.e. less biased) are the observed tetranucleotide frequencies, and thus the genomic DNA composition. Figure 3 shows how OUV varied between genomes compared to genomic GC content. Significant correlation was found between GC content and OUV values using the following regression equation:

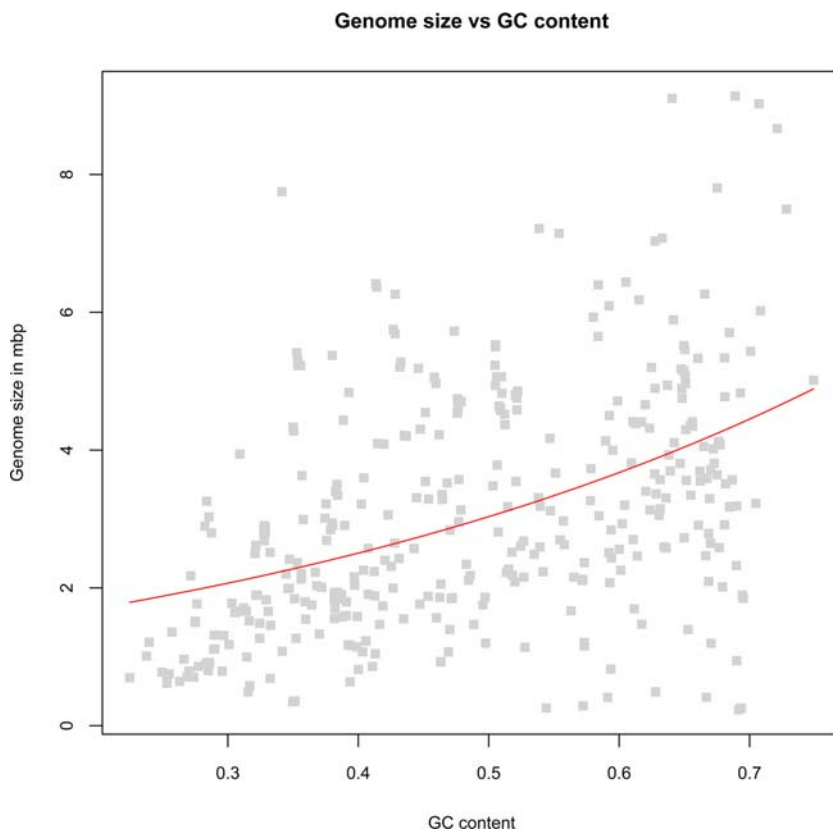
$$Y_{OUV} = \exp(-11.2 - 10.1X_{GC} + 12.4X_{GC}^2),$$

$$R^2 = 0.33, P < 0.001$$

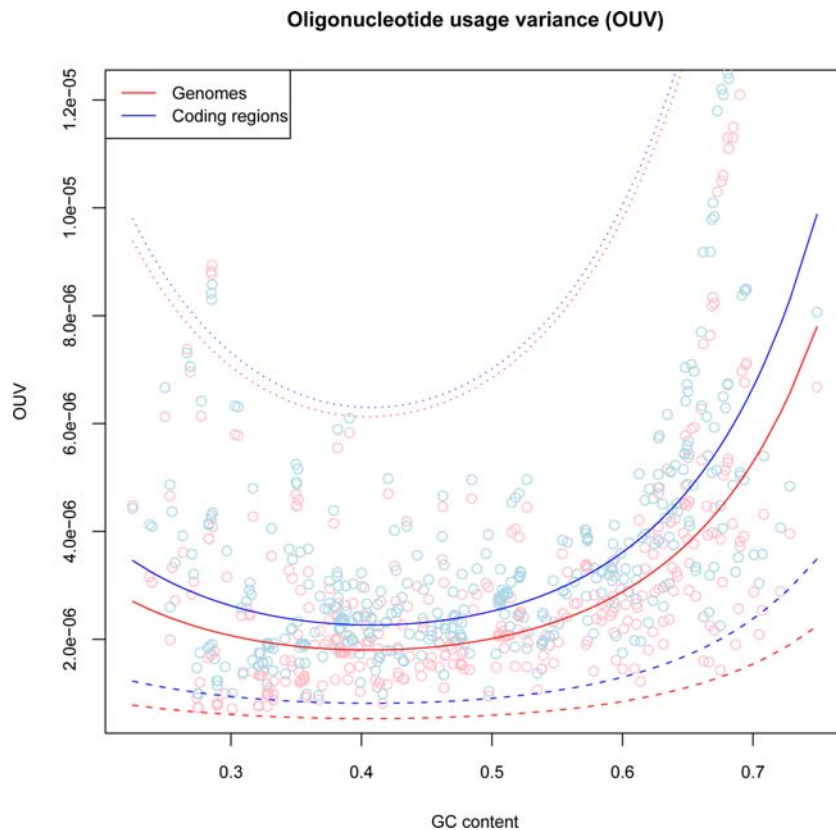
$Y_{OUV}$  designates genomic OUV values (response) while the predictor,  $X_{GC}$ , represents GC content. Our results showed that GC rich archaea and bacteria tended to have a less random DNA composition than AT rich. The reason for this is not known, but it has been argued [3] that thermodynamic properties of tetranucleotides may be important, i.e. base stacking energy and curvature. Tetranucleotide usage variance in coding regions was found to be even more strongly correlated with global GC content:

$$Y_{OUV}^c = \exp(-10.9 - 10.3X_{GC} + 12.7X_{GC}^2),$$

$$R^2 = 0.41, P < 0.001.$$



**Figure 2. Prokaryotic genome size versus GC content.** Prokaryotic chromosomes are sorted by increasing GC content from left to right on the horizontal axis. The vertical axis represents chromosome size in mbp.  
doi:10.1371/journal.pcbi.1000057.g002



**Figure 3. Tetranucleotide usage variance measures of 402 archaeal and bacterial chromosomes.** Prokaryotic chromosomes are sorted by increasing GC content from left to right (vertical axis), with red and blue regression lines representing OUV values (horizontal axis) for chromosomes and coding regions, respectively. Larger values imply more bias, or stronger selectional pressure, in genomic tetranucleotide usage. The surrounding dotted lines indicate 99% prediction intervals.  
doi:10.1371/journal.pcbi.1000057.g003

$I_{OUV}^C$  designates OUV values in coding regions (response), while  $X_{GC}$  is global GC content (predictor).

Thus, bias in tetranucleotide usage in non-coding regions was less affected by global GC content than coding regions.

Preliminary tests on a set of sequenced genomes involving randomization by increasing AT/GC content similarly to non-coding % size produced significantly larger differences in tetranucleotide usage bias than what was observed for non-coding regions. This indicates that non-coding regions do carry information and are exposed to selectional pressures and bias although considerably less than coding regions. This can be seen from Figure 3 where bias in tetranucleotide usage in coding regions increased more with GC content than tetranucleotide usage in non-coding regions. It should be noted that the above analysis is based on average values from concatenated DNA sequences, and nothing is stated about how OUV values vary within chromosomes and coding regions. Preliminary tests indicate that OUV values vary considerably within archaeal and bacterial chromosomes.

### Tetranucleotide Usage Variation Within Genomes

OUV gives a measure of how homogeneous or heterogeneous genomes are in terms of DNA composition. The OUV (but not OUD) measure can also detect to what extent tetranucleotide patterns are distributed throughout the genome. Low OUV values thus indicate increased similarity and not increased randomness. In contrast to OUV, the OUD measure is calculated as the average variance of oligonucleotide occurrences within the

chromosome based on oligonucleotide frequencies from a non-overlapping 40 kbp sliding window compared with mean genomic oligonucleotide frequencies.

From Figure 4 it can be observed that OUD scores were lower in coding regions (blue line) than in chromosomes containing both coding and non-coding regions (red line). Genomic OUD scores were progressively decreasing with increasing GC content indicating that tetranucleotide patterns in non-coding regions become progressively more similar with growing GC content. The following regression equations were obtained:

$$Y_{OUD} = \exp(-11.3 - 12.22X_{GC} + 11.1X_{GC}^2), R^2 = 0.28,$$

$$P < 0.001$$

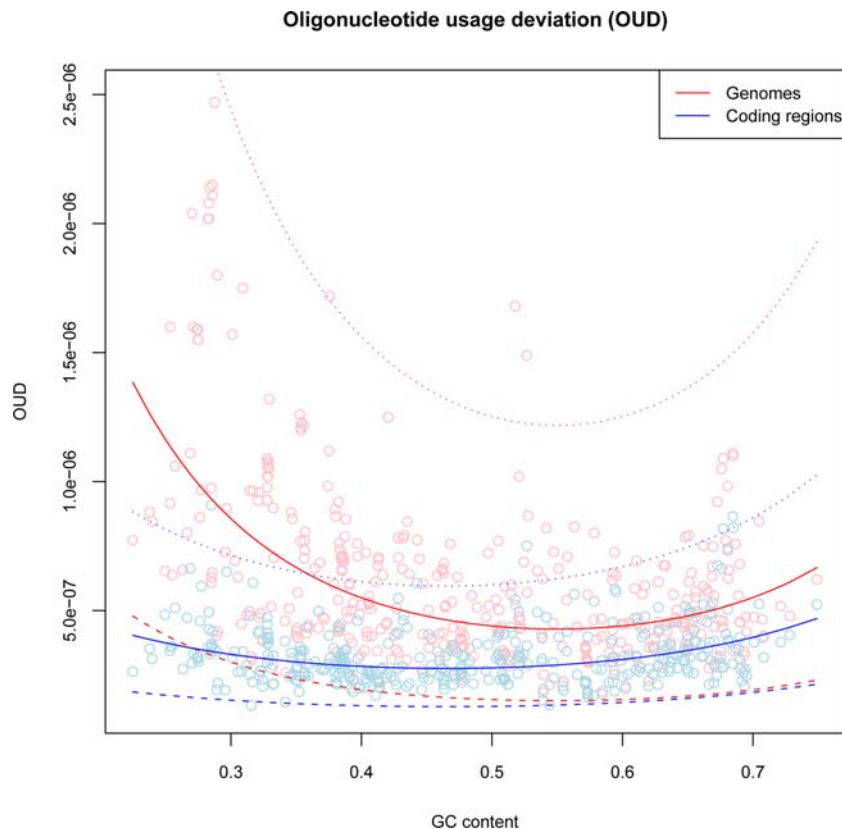
$$Y_{OUD}^C = \exp(-13.7 - 6.1X_{GC} + 6.6X_{GC}^2), R^2 = 0.11,$$

$$P < 0.001.$$

The response-functions  $Y_{OUD}$  and  $Y_{OUD}^C$  represent OUD scores in genomes and coding regions, respectively, while the predictor  $X_{GC}$  is GC content.

Difference in tetranucleotide usage within genomes was supported by a ratio test where observed tetranucleotide usage variance within genomes was divided by expected tetranucleotide variance approximated by nucleotide frequencies. From Figure 5 it can be observed that considerably less variation was detected in the coding regions (blue line) compared with chromosomes





**Figure 4. Average oligonucleotide usage deviance within prokaryotic chromosomes (OUD).** Each archaeal and bacterial chromosome is sorted on the horizontal axis with increasing GC content from left to right, with corresponding OUD values on the vertical axis. The red and blue lines represent OUD scores for whole chromosomes and coding regions, respectively. Smaller OUD values mean more homogeneous chromosomes in terms of tetranucleotide usage. The surrounding dotted lines indicate 99% prediction intervals.  
doi:10.1371/journal.pcbi.1000057.g004

containing both coding and non-coding regions (red line) with the following regression equations:

$$Y_{O_E} = \exp(0.09 + 1.44X_{GC}), R^2 = 0.38, P < 0.001$$

$$Y_{O_E}^C = \exp(0.83 + 0.55X_{GC}), R^2 = 0.07, P < 0.001$$

$Y_{O_E}$  and  $Y_{O_E}^C$  (response) are the ratios of observed divided by expected OUD values for genomes and coding regions respectively, while  $X_{GC}$  represents mean global GC content (predictor).

This variance is, however, connected to larger fluctuations in GC content within genomes in AT rich prokaryotes. See the below section on variance in GC content within genomes for more detail. No correlation was found between OUD values and genome size.

### Variation of GC Content Within Genomes

Predicted tetranucleotide usage based on genomic nucleotide frequencies was used to estimate variance in GC content within genomes (Figure 6). Since intrinsic tetranucleotide usage variance predictions were only based on nucleotide frequencies, these values were directly associated with fluctuations in local GC content obtained by comparing 40 kbp sliding windows with global (mean) GC content. We therefore wanted to investigate if fluctuations of intrinsic GC content showed any relation to global GC content, and whether there was a difference between coding and non-coding sections. Using regression analysis, significant correlation was found between global GC content and expected tetranucleotide usage variance, with the following

equation:

$$Y_{E_{OUV}} = \exp(-10.7 - 16.7X_{GC} + 14.3X_{GC}^2), \\ R^2 = 0.35, P < 0.001$$

$Y_{E_{OUV}}$  (response) represents expected OUV usage and  $X_{GC}$  GC content (predictor).

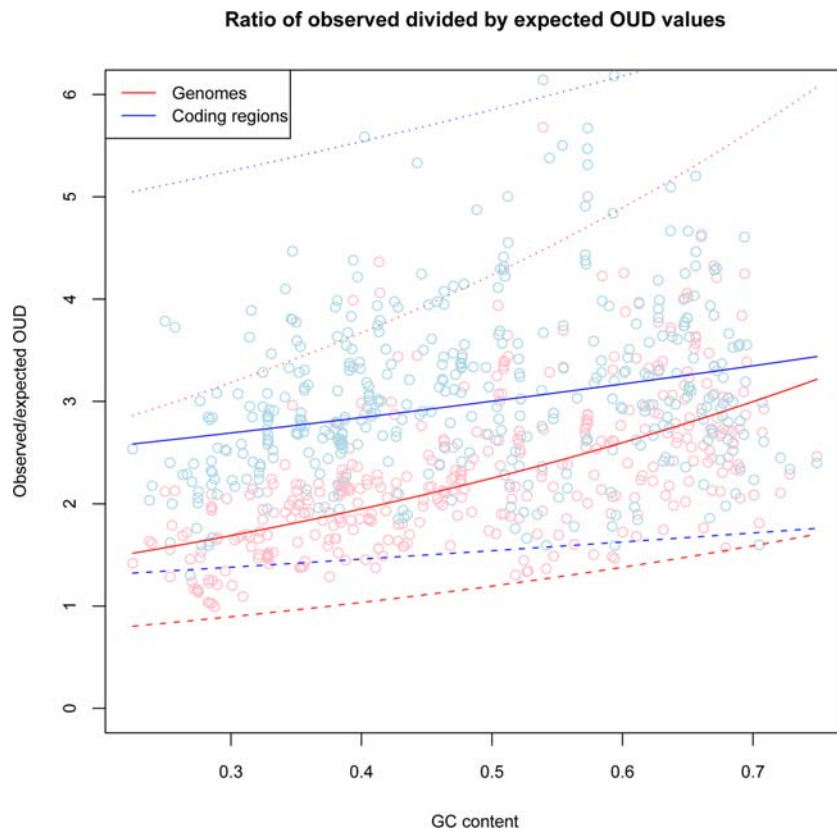
This result showed that there was significant correlation between global GC content and how GC content varied locally within genomes. The obtained correlation was also higher than what was observed for the OUD measure ( $R^2 = 0.28$ ), which means that variance in tetranucleotide frequencies within genomes is less sensitive to mean genomic GC content than variance in nucleotide frequencies.

Restricting the test to coding sections, correlation was found between global GC content and expected tetranucleotide usage variance, using the following equation:

$$Y_{E_{OUV}}^C = \exp(-13.6 - 10.6X_{GC} + 10.6X_{GC}^2), \\ R^2 = 0.10, P < 0.001$$

$Y_{E_{OUV}}^C$  represents expected OUV values in concatenated coding regions (response), and the predictor,  $X_{GC}$ , global GC content.

Thus, only weak correlation was found between fluctuations in intrinsic nucleotide frequencies and global GC content in coding regions. Since roughly 86% of prokaryotic DNA codes for



**Figure 5. Ratio of observed divided by expected OUD values.** The vertical axis shows the ratio of observed divided by expected OUD values for each chromosome sorted with respect to genomic GC content from left to right on the horizontal axis. The ratio test measures how observed oligonucleotide usage varies within chromosomes (red line) and coding regions (blue line) compared with expected based on GC content. Rising ratio values above 1 (vertical axis) means increased observed variance compared with expected. The dotted lines represent 99% prediction intervals. doi:10.1371/journal.pcbi.1000057.g005

proteins, on average, [1] the above result suggest that the GC content of non-coding regions within prokaryotes vary considerably, and that the rate of this variation within genomes is negatively correlated with mean genomic GC content. Hence, our results show that, on average, the more GC rich a genome is, the less variation of nucleotide frequencies are found in the non-coding regions, while the opposite is true for AT rich genomes, i.e. the more AT rich an archaeon or bacterium is the more varied are the nucleotide frequencies of the non-coding sections.

### Evolutionary Implications

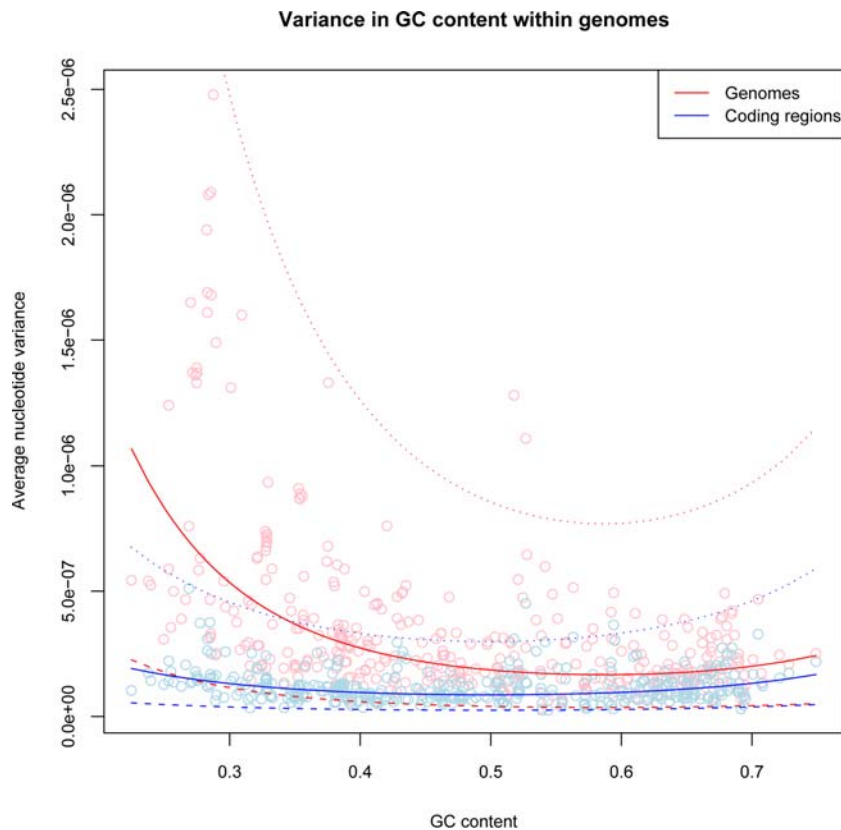
Although a link was established between genomic GC content and OUV, ( $R^2 = 0.33$ , see above section) AT rich bacteria with high OUV values and GC rich bacteria with low OUV values were also observed. High and low scores mean here being ranked among the 100 highest or lowest OUV scoring chromosomes of the 402 tested, respectively. In general, a larger number of AT rich bacteria were found with high variance scores than GC rich bacteria with low OUV scores. Closely related bacteria were also found at the opposite ends of the OUV score list. For instance, we found that “*Candidatus* Blochmannia floridanus” obtained an OUV value five times that of *Buchnera aphidicola* subsp. SG (ranked #28 and #356, respectively). Additional details can be found in Dataset S1. Both species are thought to have undergone genome reduction as they adopt a symbiotic lifestyle and have small genome size of comparable GC content [18]. Since low OUV scores imply high genomic mutation rates, different mechanisms may be responsible for the different tetranucleotide distributions in

these genomes, with *Bl. floridanus* having a more mutated genome than *B. aphidicola*. Thus, different evolutionary mechanisms may still result in similar lifestyles. Both *Bl. floridanus* and *B. aphidicola* obtained similarly high OUD scores (respectively ranked #387 and #346), which means that they both rank among the prokaryotes with the most heterogeneous chromosomes. Host associated (including pathogenic) bacteria, in general, had the most heterogeneous genomes, while free living bacteria were the most homogeneous (see Dataset S1). It should also be added that no correlation was found between OUV values and genome size, thereby removing any link between random genome composition and genome size.

### Summary

Considering prokaryotic genomes from an oligonucleotide perspective, there seemed to be little increase in information potential in oligonucleotide sizes larger than hexanucleotides.

Comparing observed to expected tetranucleotide usage (OUV), we found that coding regions are, in general, more biased than non-coding regions, and more homogeneous according to the OUD test. GC rich genomes were also found to have more biased tetranucleotide frequencies than AT rich genomes. Although AT content increased in non-coding regions in both AT rich and GC rich genomes, it did not appear to be a consequence of tetranucleotide preference. OUD was found to decrease with increasing mean genomic GC content, indicating that GC rich genomes have a more homogeneous DNA composition, especially in non-coding regions. This result was additionally supported by a



**Figure 6. Variation of GC content within genomes.** The vertical axis shows the variance of nucleotide frequencies within chromosomes (red line) and coding regions (blue line) compared with corresponding mean genomic GC content on the horizontal axis. Lower average nucleotide variance scores (vertical axis) means more similar distributions of GC content within chromosomes (and vice versa). doi:10.1371/journal.pcbi.1000057.g006

ratio test based on observed and expected OUD scores indicating that coding regions varied similarly for AT and GC rich genomes alike, while non-coding regions varied more within AT rich genomes. No correlation was found between OUV and OUD scores, implying that bias in tetranucleotide usage is not connected to intra-chromosomal homogeneity in prokaryotes.

## Materials and Methods

All sequenced archaea and bacteria available up to September 2006, (402 chromosomes and corresponding open reading frames, from 366 genomes in total) were downloaded from NCBI Genbank [19]. The different statistical tests were carried out with computer programs made according to the procedures described below. The free statistical package R [20] was used for visualization of results, regression analysis and curve-fitting. All analyses of DNA sequences were carried out in the 5'→3' direction. Analysis of genomic coding regions was performed by concatenating every open reading frame for each chromosome into one large DNA sequence.

## Notation and Formulas

For notation we let  $F_z(\cdot)$  designate frequency with respect to a DNA sequence  $z$  of an independent variable consisting of a nucleotide (AT-content,  $F_z(w)$ ,  $w$  is any nucleotide or  $F_z(A)$ ,  $F_z(G)$ ,  $F_z(C)$  and  $F_z(T)$ ) or an oligonucleotide ( $F_z(w_1w_2...w_n)$ , where  $w_1w_2...w_n$  are nucleotides making up a DNA word consisting of  $n$  nucleotides. Correlation between frequencies of DNA words was

performed with the standard Pearson correlation formula:

$$\frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_j^N (x_j - \bar{x})^2 \sum_k^N (y_k - \bar{y})^2}} \quad (1)$$

$x_i = F_x^i(w_1w_2...w_n)$  represents DNA word frequency for word  $i$  and  $\bar{x}$  is the average frequency for all possible combinations of DNA words consisting of  $n$  nucleotides (see below). The sums are taken for all possible oligonucleotide combinations, i.e.  $N = 4^n$

Average values are found using the equation:

$$\bar{x} = \frac{1}{N} \sum_i^N x_i \quad (2)$$

## Methods

The statistical information potential of the different oligonucleotide sizes was estimated by comparing frequency functions  $F_x(w_1...w_n)$  to  $F_x(w_1...w_{n-1})F_x(w_n)$  using Formula 1 for genomes with AT content between 47% and 53%. Genomes with AT content in that range were chosen due to the increased sensitivity of the Pearson correlation method for chromosomes with similar AT/GC content. By leaving one nucleotide “free” the cumulative



information in the different oligonucleotide sizes, i.e. combined information of all smaller oligonucleotides, could be measured.

GC-content in non-coding regions was calculated by finding GC-content in open reading frames and whole chromosomes with the following formula:

$$F_x^{nc}(w) = \frac{F_x^{wc}(w) - cF_x^c(w)}{nc} \quad (3)$$

where  $c$  = coding fraction measured by the sum of open reading frame sizes divided by corresponding chromosome size. Non-coding fraction is then  $nc = 1 - c$ .

The superscripts  $c$ ,  $nc$ ,  $wc$  designate GC content in coding, non-coding and whole chromosomes, respectively.

GLM based regression analysis was used to examine the relationship between global GC content (predictor) and genome size (response). Because of the highly non-linear nature of the data, Spearman's rank based correlation test was additionally used between genome size and global GC content.

OUV was calculated for each genome with the following formula:

$$\sum_i^N (x_i - E(x_i))^2 \quad (4)$$

where  $E$  is expected frequency for DNA word  $x_i$  based on nucleotide frequencies:

$$E(x_i) = F_x(w_1)F_x(w_2) \dots F_x(w_n). \quad (5)$$

The expected value is thus calculated based on the assumption that each oligonucleotide frequency consists of individual nucleotide frequencies independent of each other. Since Formula 5 bases calculations of the expected value only on genomic nucleotide frequencies it can be considered as a measure of how "random" the genomic composition of the organism is. Low OUV scores can therefore be interpreted to mean that a high degree of non-directed mutations has taken place within the genome, or, alternatively, that the individual nucleotides in each tetranucleotide are more 'loosely' tied to each other. High OUV scores may also be taken to mean that a larger degree of bias, or order, is present in the genomic oligonucleotide usage. An additional point of view is that OUV is a measure of information potential, where low OUV values mean that the DNA sequence in question has lower statistical potential of carrying information (and vice versa).

OUV was calculated for both whole chromosomes and coding regions. Regression analysis was then carried out between OUV values (response), both coding and whole chromosomes, and global GC content (predictor). The resulting regression equations, with corresponding coefficient of determination, here denoted by  $R^2$ , can be found in the Results/Discussion section together with the corresponding  $P$  values.

The OUD test gives an average estimate of how oligonucleotide frequencies vary within prokaryotic chromosomes. Variance is calculated between the oligonucleotide frequencies calculated from a 40 kbp non-overlapping sliding window and mean genomic oligonucleotide frequencies with the following formula:

$$\frac{1}{M} \sum_j^M \sum_i^N (z_i^j - \mu_x)^2. \quad (6)$$

Summations are taken over all possible oligonucleotide combina-

tions  $1 \leq i \leq N$ , and all non-overlapping sliding windows  $1 \leq j \leq M$  in DNA sequence  $x$ .  $z_i^j = F_x^i(w_1 w_2 \dots w_n)$  represents the  $n$ -word frequencies for sliding window  $j$ , while  $\mu_x = F_x(w_1 \dots w_n)$  is the mean frequency of word  $w_1 \dots w_n$  in DNA sequence  $x$ .

Regression analysis was performed for both whole chromosomes and coding regions with OUD as response and global GC content as predictor. The resulting regression equations with corresponding coefficient of determination,  $R^2$ , and  $P$ -values can be found in the Results/Discussion section.

The ratio of observed divided by expected OUD values was additionally used to test whether any fluctuations in tetranucleotide frequencies could be detected in coding and non-coding regions. To calculate this, the OUD values obtained for each chromosome with Formula 6 was divided by the following equation:

$$\frac{1}{M} \sum_j^M \sum_i^N (E(z_i^j) - E(x_i))^2, \quad (7)$$

which resulted in the formula:

$$\sum_j^M \sum_i^N \left( \frac{z_i^j - \mu_x}{E(z_i^j) - E(x_i)} \right)^2. \quad (8)$$

In this case, the variance was calculated, for each chromosome, between expected oligonucleotide frequencies from a 40 kbp sliding window and expected oligonucleotide frequencies based on Formula 5.

Variation of nucleotide frequencies within genomes was calculated for each chromosome similarly to the OUD test, but expected oligonucleotide frequencies based on Formula 5 were used instead of observed. This is the same as calculating the variance between local and global GC content (i.e. GC content) and Formula 7 was used for this as well.

Regression analysis was performed on values obtained with Formula 7 and global GC content for both coding and whole chromosomes. The resulting equations, with the corresponding coefficients of determination,  $R^2$ , and significance, can be found in the Results/Discussion section.

## Supporting Information

**Dataset S1** Microsoft Excel file consisting of the data used to generate the results in the manuscript. Each column is labeled according to the abbreviations used in the text and additionally explained on a separate sheet.

Found at: doi:10.1371/journal.pcbi.1000057.s001 (0.17 MB XLS)

## Acknowledgments

Special thanks to Simon Hardy and the reviewers for constructive comments and for critically reading through the manuscript. Also, thanks to Stig Larsen for help with the statistical analysis and to Stein Marvold for the graphical image.

## Author Contributions

Conceived and designed the experiments: JB. Performed the experiments: JB. Analyzed the data: JB ES DU. Contributed reagents/materials/analysis tools: JB. Wrote the paper: JB ES DU. Revised and drafted the manuscript: DU.

## References

1. Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A* 101: 3160–3165.
2. Sinden RR (1994) *DNA structure and function*. San Diego: Academic Press. pp 398.
3. Reva ON, Tummler B (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* 5: 90.
4. Packer MJ, Dauncey MP, Hunter CA (2000) Sequence-dependent DNA structure: dinucleotide conformational maps. *J Mol Biol* 295: 71–83.
5. Packer MJ, Dauncey MP, Hunter CA (2000) Sequence-dependent DNA structure: tetranucleotide conformational maps. *J Mol Biol* 295: 85–103.
6. Ussery D, Soumpasis DM, Brunak S, Staerfeldt HH, Worning P, Krogh A (2002) Bias of purine stretches in sequenced chromosomes. *Comput Chem* 26: 531–541.
7. Kulkarni OC, Vigneshwar R, Jayaraman VK, Kulkarni BD (2005) Identification of coding and non-coding sequences using local Holder exponent formalism. *Bioinformatics* 21: 3818–3823.
8. Allen TE, Price ND, Joyce AR, Palsson BO (2006) Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. *PLoS Comput Biol* 2: e2. doi:10.1371/journal.pcbi.0020002.
9. Chen LL, Zhang CT (2003) Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. *Biochem Biophys Res Commun* 306: 310–317.
10. Foerster KU, von MC, Hooper SD, Bork P (2005) Environments shape the nucleotide composition of genomes. *EMBO Rep* 6: 1208–1213.
11. Willenbrock H, Friis C, Juncker AS, Ussery DW (2006) An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biol* 7: R114.
12. Rocha EP (2004) The replication-related organization of bacterial genomes. *Microbiology* 150: 1609–1627.
13. Worning P, Jensen LJ, Hallin PF, Staerfeldt HH, Ussery DW (2006) Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* 8: 353–361.
14. Karlin S, Mrazek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* 179: 3899–3913.
15. Kozobay-Avraham L, Hosid S, Bolshoy A (2006) Involvement of DNA curvature in intergenic regions of prokaryotes. *Nucleic Acids Res* 34: 2316–2327.
16. Musto H, Naya H, Zavala A, Romero H, varez-Valin F, Bernardi G (2006) Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochem Biophys Res Commun* 347: 1–3.
17. Musto H, Naya H, Zavala A, Romero H, varez-Valin F, Bernardi G (2005) The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: a reply to Marashi and Ghalanbor. *Biochem Biophys Res Commun* 330: 357–360.
18. Gil R, Silva FJ, Zientz E, Delmotte F, Gonzalez-Candelas F, et al. (2003) The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc Natl Acad Sci U S A* 100: 9388–9393.
19. National Center for Biotechnology Information (2007) Genbank database. ;Available at: <http://www.ncbi.nlm.nih.gov/Genomes/>. Accessed 15 March 2008.
20. R Development Core Team (2007) R: A language and environment for statistical computing, version 2.5.1 [computer program]. ;Available at: [www.r-project.org](http://www.r-project.org). Accessed 15 March 2008.